Application of Artificial Neural Networks for Detecting Diabetes Mellitus Using Demographic, Clinical, Lifestyle, and Dietary Risk Factors: A Case Study from Kaura Namoda, Nigeria

Abstract

Diabetes Mellitus (DM) is a chronic condition that demands urgent attention due to its widespread prevalence and severe complications. Early detection of DM at the grassroots level through patient risk factors is critical for effective prevention and management. This study utilized data from 400 patients collected from medical records at General Hospital Kaura Namoda, Zamfara State, Nigeria, spanning January 2019 to December 2023. The data was divided into two sets: the first set included demographic, clinical, and lifestyle risk factors, while the second set also incorporated dietary risk factors. Analysis was conducted using R statistical software (version 3.13), and the results demonstrated the effectiveness of the Multi-Layer Perceptron Neural Network (MLPNN) model. For the first dataset, the MLPNN model achieved detection rates of 97.5% for DM patients in the training sample, 94.9% in the validation sample, and 88.9% in the test sample. Similarly, non-DM detection rates were 94.9% in the training sample, 82.6% in the validation sample, and 84.6% in the test sample. For the second dataset, which included dietary risk factors, the model performed even better, achieving 99.2% detection for DM patients in the training sample, 100% in the validation sample, and 100% in the test sample. Non-DM detection rates were 98.7% in the training sample, 95.7% in the validation sample, and 100% in the test sample. The study concludes that incorporating dietary risk factors alongside demographic, clinical, and lifestyle factors significantly improves the accuracy of DM and non-DM detection, underscoring their importance in predictive modeling.

Keywords: Diabetes Mellitus, Backpropagation, Detection, Artificial Neural Network, Receiver Operating Characteristic Curve.

1. Introduction

Diabetes Mellitus (DM) is a chronic disorder characterized by abnormally high blood sugar levels (glucose). In people with DM, blood sugar levels remain high. This might be because insulin is not being produced at all, is not made at a sufficient level or is not as effective as it should be. DM affects more than 300 million people worldwide. In 2016, it was discovered that 1 in 5 people aged 50 years and above has DM. The highest prevalence (17.9%) was among American Indians and Alaska natives. DM cases are increasing worldwide, and countries are struggling to fight the disease (WHO, 2004). The misconception that DM is "a disease of the wealthy" is still held by some people. Still, the evidence published in the Diabetes Atlas of the International Diabetes Federation (IDF, 2021) disproves that delusion: 80% of people with DM live in low and middle-income countries, and socially disadvantaged countries are the most vulnerable to that disease. Today's emerging DM hotspots include countries in the Middle East, Europe, Western Pacific and South-East Asia, where economic development has transformed lifestyles. These rapid transitions are bringing previously unheard rates of obesity and DM; developing countries are facing a firestorm of ill health with inadequate resources to protect their population. Thus, it is necessary to increase awareness of the importance of a healthful diet and physical activity, especially for children and adolescents. Conducive environments have to be created that lay the foundations for healthy living (NIH, 2021).

Nigeria has the largest population in Africa, roughly about 220 million; of this, the adult population aged 20–79 years is approximately 140 million. One-third of all the cases of DM are in rural communities, while the rest are in the urban centres. About 5 million of the cases of DM in Nigeria are undiagnosed, deaths related to DM in Nigeria in 2023 were estimated to be 2I5, 137, and the current prevalence of DM in Nigeria is roughly from 8% to 10%. Of the four classes of DM, two types are frequently found in Nigeria: type 1 DM and type 2 DM. Also, among the two, type 2 DM is the most common and accounts for about 90% to 95% of all cases of DM. The prevalence of type 1 DM is unknown, but there are few reports from various parts of Nigeria. Its prevalence ranges from 0.1/1000 to 3.1/ 10000, and 1 out of every 17 adults have the disease, National Institute of Health (NIH, 2021). Moreover, the pooled prevalence of DM in the six geopolitical zones of Nigeria was 3.0% in the North- West, 5.9 in the North-East, 3.8% in the North- Central, 5.5% in the South-West, 4.6 % in the South – East and 9.8% in the South-South (NIH, 2021).

Today, many techniques have been developed for data mining, and there is an art to selecting and applying the best method for a particular situation. Methods for analyzing data can be divided into two groups: supervised learning and unsupervised learning. Supervised learning requires input data that has both independent variables or input variables and a dependent variable or output variable whose value is to be estimated. By various means, the process learns how to predict the value of the output variable based on the input variables. Decision Trees (DT), Regression Analysis (RA) and Artificial Neural Networks (ANNs) are examples of supervised learning. Unsupervised learning does not identify output variables but rather treats all of the variables equally. In this case, the goal is not to predict a variable's value but to look for patterns, groupings or other ways to characterize the data that may lead to understanding how the data interrelates. Cluster Analysis (CA), Correlation, Factor Analysis (FA), Principal Components Analysis (PCA) and statistical measures are examples of unsupervised learning (Bellazi and Zupan, 2008; Al-Shaye, 2011).

Bellazi and Zupan (2008) Artificial Neural Networks are popular data mining tools for building complicated models. An Artificial Neural Network Model contains three layers: an input layer, an intermediate hidden layer and an output layer. Also, each layer comprises nodes (neurons) and links. The nodes in the input layer are viewed as predicted variables, whereas the nodes in the output layer are analyzed as the outcome variables. The paper used a popular ANN Architecture called Multilayer Perceptron Neural Network (MLPNN) with back-propagation (i.e. Supervised Learning Algorithm), arguably the most commonly used and well–studied ANN architecture. MLPNN is feed- a feed-forward neural network trained with the standard back-propagation algorithm, and they are known to be a powerful function approximator for prediction and classification problems (Xue-Hui Meng *et al.,* 2011). Artificial Neural Network provides a general way of approaching problems. When the network's output is categorical, it performs prediction, and when the output has discrete values, it does classification (Al-Shaye, 2011). The paper reviewed work on ANN for the prediction of Diabetes, such as Sahu and Mantri (2023) used the MLPNN model for the prediction of Diabetes using demographic and clinical risk factors in the face of inconsistent results, gaps and data class imbalance.
The Model achieved a prediction accuracy of 84% relative to the baseline.

Similarly, Chen *et al.* (2024) observed that ANNs trained using risk factors had better efficacy and facilitated the reduction of harm caused by type 2 DM combined with Hyperuricaemia. Likewise, Bukhari *et al*. (2021) used demographic, clinical and lifestyle risk factors to train the Artificial Backpropagation Stochastic Gradient Neural Network (ABPSCGNN) algorithm for the prediction of Diabetes patients; the ABPSCGNN model achieved 93 prediction accuracy. Also, Pradhan *et al.* (2020) also applied the MLPNN model for predicting Diabetes patients using nine (9) features. The Model had 85.09% prediction accuracy. Moreover, Setiawan *et al.* (2024) focused on the Neural Network model for predicting Diabetes patients using clinical data. The result obtained showed that the Model had an accuracy of 97%, demonstrating the Model's ability to predict diabetes patients. However, all the work reviewed used clinical risk factors, demographic and clinical risk factors or combined demographic, clinical and lifestyle risk factors. Still, there is a need to include dietary risk factors for better prediction accuracy.

Materials and method

The sample data (400 patients) used in this paper was obtained from patients' records suffering from DM in General Hospital Kaura Namoda Local Government Zamfara State, Nigeria, from January 2019 to December 2023. The sample data also consists of 14 risk factors: 2 demographic risk factors (age and sex), four clinical risk factors (family history of DM, blood glucose level, blood pressure level and body mass index), one lifestyle risk factor (physical activity), 7 dietary risk factors (preference for sweet food, preference for salty food, red meat, refined carbs, energy drinks, white rice and processed meats) and 1 output (i.e. diagnosis recommended for these patients by the physician that attended to them. Similarly, the sample data was divided into two sets; the first set consisted of 400 patients with demographic, clinical and lifestyle risk factors, while the second set consisted of 400 patients with demographic, clinical, lifestyle and dietary risk factors. The data risk factors and their formats are presented in Table 1.

Variable Name	**Classification Network Type**	**Predictive Network Type**
14 risk factors	Y or N (Character)	1 or 0 (Continuous)

Diagnosis	DM Non- DM	1 0

The feature selection method was used to identify risk factors that are not useful and do not contribute significantly to the neural network's performance. In this regard, the backwards stepwise method was used for risk factor selection, and the risk factors selected were the patient's age, family history of DM, blood glucose level, blood pressure, body mass index, physical activity, preference for sweet food, red meats, refined carbs, energy drinks, white rice and processed meat and those that were not selected are sex and preference for salty food. The selected risk factors (12) were used to train the MLPNN algorithm, and the Model obtained was used to detect DM and non-DM patients.

2. Design of Multilayer Perceptron Neural Network

A three-layer network was used for the design of MLPNN. That is, the MLPNN design has 12 input layers, 24 hidden layers and 1 output layer. Figure 1 shows a diagrammatical representation of the proposed Neural Network (NN). To search for the optimal training parameter, the learning rate, the momentum rate, the transfer function in the hidden layer and the learning algorithm back propagation (BP) were implemented. The idea of the BP was to reduce error until the network learns the training data. The training starts with random weights, and the aim is to adjust them so that the learning error is very small. The network nodes in the BP algorithm are arranged in layers so that they can send their signal forward, and then the learning error is calculated and propagated backwards until it meets a satisfactory learning error. During the course of training the Model, the network was monitored so that it corrected itself in order to achieve the minimum possible error. The training dataset is presented to the input layers, and the network propagates the input pattern from layer to layer until the output pattern is generated. Then, output was

obtained from a summation of the weighted input of a node and mapped to the network activation function.

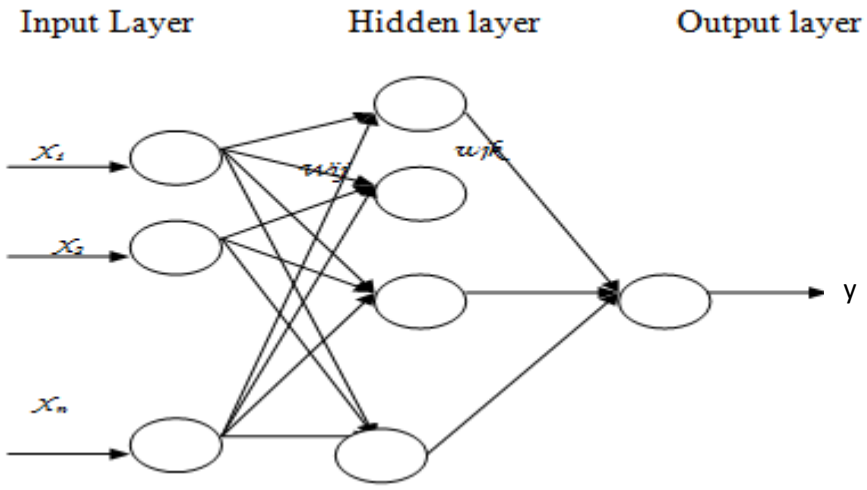

Figure 1: Design of Multilayer Perceptron Neural Network

Equations (2.1) and (2.2) showed the calculation formula from input layer (*i*) to the hidden layer (*j*), where O_j is the output of node j, O_i is the output of node I, w_{ij} is the weight connected between node i and node j, and θ_j is the bias of node j.

$$O_j = f\left(net_j\right) = \frac{1}{1+e^{-net_j}} \tag{2.1}$$

$$net_j = \sum_i w_{ij} O_i + \theta_j \tag{2.2}$$

Similarly, Equation (2.3) and (2.4) showed computation formula for hidden layer (*j*) to output layer (*k*), where O_k is the output of node k, O_j is the output of node j, w_{jk} is the weight connected between node j and k, and θ_k is the bias of node k.

$$O_k = f\left(net_k\right) = \frac{1}{1+e^{-net_k}} \tag{2.3}$$

$$et_k = \sum_k w_{jk} O_j + \theta_k \tag{2.4}$$

The network activation function in Equations (2.1) and (2.3) was Sigmoid Activation Function. Moreover, error is calculated using Equation (2.5) to measure the differences between desired

output and actual output that had been produced in feed forward phase. Error was then propagated backward through the network from output layer to input layer and weights are modified to reduce the error as the error was propagated.

$$Error = \frac{1}{2}\left[Output_{desired} - Output_{actual}\right]^2 \tag{2.5}$$

Based on the error calculated, backpropagation was applied from output (k) to hidden (j) as in Equations (2.6) and (2.8)

$$w_{ji}\left(t+1\right) = w_{ji}\left(t\right) + \Delta w_{ji}\left(t+1\right) \tag{2.6}$$

$$\Delta w_{ji}\left(t+1\right) = \eta\delta_k O_j + \alpha\Delta w_{ji}\left(t\right) \tag{2.7}$$

$$\delta_k = O_k\left(1-O_k\right)\left(t_k - O_k\right) \tag{2.8}$$

where $w_{ji}\left(t\right)$ is the weight from node j to node i at time t, Δw_{ji} is the weight adjustment, η is the learning rate, α is the momentum rate, δ_j is an error at node j, δ_k is an error at node k, O_i is the actual network output at node i, O_j is the actual network output at node j, O_k is the actual network output at node k, w_{kj} is the weight connected between node j and k, and θ_k is the bias of node k. This process was repeated iteratively until convergence was achieved (targeted learning error).

3. Discussions of Results

For the first set of data, which consists of 400 patients with demographic, clinical and lifestyle risk factors, the trained MLPNN model was used to detect DM and Non-DM patients in the training, validation and test samples. Table 1 shows the results, which indicated that the MLPNN model detected 97.5% of patients with DM in the training sample, 94.9% in the validation sample and 88.9% in the test sample. Similarly, the Model also detected 94.9% of patients as non-DM in the training sample, 82.6% in the validation sample and 84.6% in the test sample.

Table 1: Detection of DM and Non-DM Patients using MLPNN Model

Observed		Detected Patients			
		DM	Non-DM	Total	Per cent Correct
Training Sample	DM	*119*	*3*	*122*	*97.5*
	Non-DM	*8*	*150*	*158*	*94.9*
	Total	127	153	280	
Validation Sample	DM	*54*	*3*	*57*	*94.7*
	Non-DM	4	*19*	*23*	*82.6*
	Total	*58*	*22*	80	
Test Sample	DM	*24*	*3*	27	88.9
	Non-DM	*2*	*11*	13	84.6
	Total	*26*	*14*	40	

Similarly, Table 1 also revealed that out of 122 DM patients in the training sample, the Model detected 119 patients with DM and 3 patients as non-DM. In the validation sample, out of 57 patients, the Model detected 54 patients with DM and 3 patients as Non-DM, and in the test sample, out of 27 patients, the Model detected 24 patients with DM and 3 patients as Non-DM. Likewise, out of 158 Non-DM patients in the training sample, the Model detected 150 patients as Non-DM and 8 patients with DM; in the validation sample, out of 23 Non-DM patients, the Model detected 19 patients as Non-DM and 4 patients with DM and in the test sample out of 13patients the Model detected 11 patients as Non-DM and 2 patients with DM. In the second set of data, which consists of 400 patients with demographic, clinical, lifestyle and dietary risk factors, Table 2 showed that the MLPNN model detected 99.2% of patients with DM in the training sample, 100% in the validation sample and 100% for the test sample. Also, the Model detected 98.7% of patients as non-DM in the training sample, 95.7% in the validation and 100% in the test sample.

Table 2: Detection of DM and Non-DM Patients using MLPNN Model

Observed		Detected Patients			
		DM	Non-DM	Total	Per cent Correct
Training Sample	DM	*121*	*1*	*122*	*99.2*
	Non-DM	*2*	*156*	*158*	*98.7*
	Total	123	157	280	
Validation Sample	DM	*57*	*0*	*57*	*100.0*
	Non-DM	1	*22*	*23*	*95.7*
	Total	*58*	*22*	80	
Test Sample	DM	*27*	*0*	27	100.0
	Non-DM	*0*	*13*	13	100.0
	Total	*27*	*13*	40	

Similarly, Table 2 also revealed that out of 122 patients with DM in the training sample, the Model detected 121 patients with DM and 1 patient as Non-DM. In the validation sample, out of 57 patients, the Model detected 57 patients with DM and no patients as Non-DM, and in the test sample, out of 27 patients, the Model detected 27 patients with DM and no patients as Non-DM. Likewise, out of 158 Non-DM patients in the training sample, the Model detected 156 as Non-DM patients and 2 patients as DM; in the validation sample, out of 23 Non-DM patients, the Model detected 22 as Non-DM patients and 1 patient with DM and in the test sample out of 13 patients the Model detected 13 as Non-DM patients and no patients with DM.

MLPNN Model was evaluated in terms of its accuracy, sensitivity, and specificity in the detection of DM and non-DM. The results obtained in the training, validation and test samples for the first and second data sets are presented in Tables 3 and 4, respectively. The paper found that for the first set of data, in the training sample, the MLPNN Model achieved an accuracy of 96.1% with a sensitivity of 97.5% and a specificity of 94.9%. In the validation sample, the

Model achieved an accuracy of 91.3% with a sensitivity of 94.7% and a specificity of 82.6%, and in the test sample, the Model showed an accuracy, sensitivity and specificity of 87.5%, 88.9% and 84.6%, respectively.

Table 3: Evaluation of Model Performance for First Set of Data

Indices	Training Sample	Validation Sample	Test Sample
Accuracy (%)	96.1	91.3	87.5
Sensitivity (%)	97.5	94.7	88.9
Specificity (%)	94.9	82.6	84.6

For the second set of data, in the training sample MLPNN Model achieved accuracy of 98.9% with sensitivity of 99.2% and specificity of 98.7%. In the validation sample the Model achieved accuracy of 98.8% with sensitivity of 100% and specificity of 95.7% and in the test sample the Model achieved accuracy of 100% with sensitivity of 100% and specificity of 84.6%.

Table 4: Evaluation of Model Performance for Second Set of Data

Indices	Training Sample	Validation Sample	Test Sample
Accuracy (%)	98.9	98.8	100
Sensitivity (%)	99.2	100	100
Specificity (%)	98.7	95.7	100

The Receiver Operating Characteristic (ROC) Curve was used to determine the discriminatory ability of the MLPNN model in distinguishing between DM and Non- DM patients. This was done by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) at different cut-off points (Zweigh and Campbell, 1993). Similarly, the paper used the interpretation given by the Traditional Academic Point System (TAPS, 2005) to interpret the Area under the Receiver Operating Characteristic (AUROC) Curve of the Model, such as Area less than equal to 0.59 indicate poor discrimination, 0.60 to 0.69 fair dis ion, 0.70 to 0.79

good discrimination, 0.80 to 0.89 very good discrimination and 0.90 to 1.00 excellent discrimination.

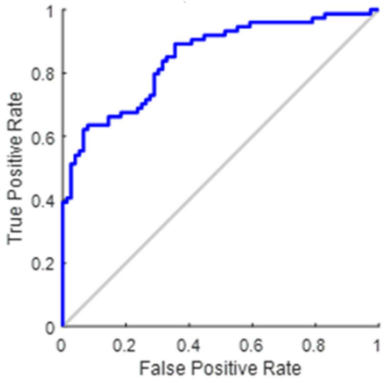

Figure 2: ROC Plot of First Set of Data

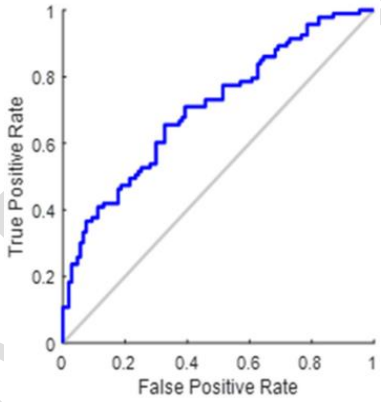

Figure 3: ROC Plot of Second Set of Data

Figures 2 and 3 showed ROC Curves for the first and second data sets. The MLPNN Model had an AUROC Curve of 0.96 for the first set of data and 0.99 for the second set of data. The two AUROC curves fall within 0.90 to 1.00; this indicates excellent discrimination, and the Model can discriminate between DM and Non-DM patients. However, in the second set of data that used demographic, clinical, lifestyle and dietary risk factors, the AUROC Curve was larger than the AUROC Curve of the first set of data that used demographic, clinical and lifestyle risk factors only, and this was attributed to the inclusion of dietary risk factors in the second set of data.

4. Conclusion

Diabetes Mellitus (DM) is a chronic disorder which needs urgent attention. Detection of DM from the grass root using patient risk factors is the key to early prevention of the disease. The sample data of 400 patients used in this paper was obtained from past patient records of patients suffering from DM in General Hospital Kaura Namoda, Zamfara State, from January 2019 to December 2023. The data was divided into two sets. The first set consisted of 400 patients with demographic, clinical and lifestyle risk factors, and the second set consisted of 400 patients with demographic, clinical, lifestyle and dietary risk factors. MLPNN algorithm was trained using the significant risk factors, which were selected using the feature selection method. The result indicated that for the first set of data, the MLPNN model detected 97.5% of patients with DM in the training sample, 94.9% in the validation sample and 88.9% in the test sample. Similarly, the Model detected 94.9% of patients as non-DM in the training sample, 82.6% in the validation sample and 84.6% in the test sample. For the second set of data, the Model detected 99.2% of patients with DM in the training sample, 100% in the validation sample and 100% in the test sample. Also, the Model detected 98.7% of patients as non-DM in the training sample, 95.7% in the validation and 100% in the test sample. The MLPNN model had an AUROC Curve of 0.96 for the first set of data and 0.99 for the second set of data. The two AUROC curves fall within 0.90 to 1.00; this indicates excellent discrimination, and the Model has the ability to discriminate between DM and Non-DM patients. Evaluation of the model performance revealed that for the first set of data, in the training sample, the MLPNN model achieved an accuracy of 96.1% with a sensitivity of 97.5% and a specificity of 94.9%. In the validation sample, the Model achieved an accuracy of 91.3% with a sensitivity of 87.5% and a specificity of 82.6%, and in the test sample,

the Model showed an accuracy of 87.5 with a sensitivity of 88.9% and a specificity of 84.6%. Likewise, for the second set of data in the training sample, the Model achieved an accuracy of 98.9% with a sensitivity of 99.2% and specificity of 98.7%; in the validation sample, the Model achieved an accuracy of 98.8% with a sensitivity of 100% and specificity of 95.7% and in the test sample the Model achieved an accuracy of 100% with sensitivity of 100% and specificity of 100%. The paper concludes that the use of demographic, clinical, lifestyle and dietary risk factors increases the accuracy, AUROC Curve, sensitivity and specificity of the Model in the detection of DM and Non-DM patients.

References

Al-Shayea, Q.K. (2011). Artificial Neural Network in Medical Diagnosis. *International Journal of Computer Science* 8(2):150-154.

Bellazi, R. and Zupan, B. (2011). Predictive Data Mining in Clinical Medicine: Current Issues and Guidelines. *International Journal of Medical Information* 8 (77):81-97.

Bukhari, M., Alkhamees, B.F., Hussain, S., Gumaei, A., Assiri, A. and Ullah, S.S. (2021). An Improved Artificial Neural Network for Effective Diabetes Prediction. *Hindawi Complexity*, doi.org/10.1156/2021/552527/.

Chen, Q., Hu, H., She, Y., He, Q., Huang, X., Shi, H., Cao, X. and Zhang, X. (2024). An Artificial Neural Network Model for Evaluating the Risk of Hyperuricaemia in Type 2 Diabetes Mellitus. *Science Report*, 14: doi.org/10.1038/s41598-024-52550-1.

International Diabetes Federation (2021). Diabetes Prevalence in 2021. Retrieved from diabetesatlas.org

National Institute of Health (2021). Diabetes Mellitus. Retrieved from https://www.ncbi.nlm.nih.gov www.ncbi.nlm.gov

Pradhan, N., Rani, G., Dhaka, V.S. and Poonia, R.C. (2020). Diabetes Prediction using Artificial Neural Network. *Journal of Science Direct:* doi.org/10.1016/8978-0-12-819061-6.00014-8.

Rackel, R.E. (2007). *Textbook of Family Medicine*. Seventh edition, France: Saunders Elsevier Publishers. Pages 989-1000.

Sahu, P. and Mantri, J.K. (2023). Artificial Neural Network Based Diabetes Prediction Model and Reducing Impact of Class Imbalance on its Performance. *Journal of SSRN*, 15: doi.org/10.2139/ssrm.4538967.

Setiawan, H. (2024). Enhancing the Accuracy of Diabetes Prediction Using Feed Forward Multilayer Perceptron Neural Networks. Journal *of Brilliance Research of Artificial Intelligence*, 4(1): doi.org/1047709/brilliance.v4i1.3888.

Traditional Academic Point System (2005). Interpretation of Area Under the Receiver Operating Characteristic Curve. *Journal of Maths Psychology* 18: 156-189.

World Health Organization (2004). Action Plan for the Global Strategy of Prevention and Control of Non-Communicable Disease. *World Health Organization Journal* 15(4):17-19

Xue- Huimeng, G., Yi- Xiang, H., Dong-Ping, R., Qiu -Zhang, H. and Qing- Liu, K.(2013). Comparison of Three Data Mining Models for Predicting Diabetes or Pre Diabetes by Risk Factors. *Kaohsiung Journal of Medical Sciences* 29: 93-99.

Zweig, P. and Campbell, G. (1993). Advances in Statistical Methodology for Evaluation of Diagnostic and Laboratory Tests. *Journal of Medical Science*, 13: 499-508.